[專案一]

寶貝的名字

美國社保局 (Social Security Administration) 保有當年度出生嬰兒最流行取的名字排行榜 (見：social security baby names)。

請瀏覽一下附件檔案 baby1990.html, baby1992.html ... ，內含近似於社保局網頁中的 html 原始碼，同時思考如何從中爬取一些資料，從而寫入你的程式碼到 childnames.py 中。

Part A

➡ [Part B]

1) 撰寫程式讓使用者在命令列下指令爬取資料。指令格式如下：

   `usage: [--summaryfile] file [file ...]`

   ** 方括號內為選項，其它則為必要引數 (在此專案為打算要解析的檔案名稱)。若使用者輸入錯誤，應提示正確格式 (舉例如下)。

```
C:> python childnames.py
usage: [--summaryfile] file [file ...]

C:> python childnames.py baby2006.html
2006
Aaliyah 91
Aaron 57
...
```

2) 請在 childnames.py 中撰寫函數：`extract_names(filename)`，用來讀取檔案 (如：`baby2006.html`)，並傳回一個串列，內含：第一個元素為 **年度** 字串，接著的是該年度的 **名字-排行** 字串，依字母排序。如：`['2006', 'Aaliyah 91', 'Abagail 895', 'Aaron 57', ...]`。

3) 同時請修改函數 `main()` 以呼叫你剛完成的 `extract_names()` 函數，並印出它所傳回的資料。(專案中 `main()` 函數內已經有提供些程式碼用以解析命令列

引數) 。

** 請留意一般來說，正則表達式 (regular expressions) 不是解析網頁的好工具，不過此專案中的網頁比較簡單而且格式一致。

4) 無需分開處理男孩與女孩的名字，只需把它們放在一起即可。某些年份的 html 資料中的部份名字不只出現一次;但我們每個名字只取一個數字(排行)，或者你也可以使用聰明的演算法選擇其中最小的數字。

** 學習老鳥寫程式的訣竅：制定程式撰寫的進程，每一階段都可執行並輸出一些資料，再進行下一階段；而不是嘗試一次就要完成整個專案。

每個階段結束時印出資料可以幫你思考如何重構這些資料供下一階段使用。Python 很適合這種漸進式開發的風格。

5) 以下是建議的開發里程碑：

- 讀取檔案並印出 所有文字 資料
- 搜尋、擷取 年份 (year) 資料並印出
- 擷取 名字與排行 (name-rank) 資料並列印
- 建立 字典，存放名字並列印
- 建立 串列 [year, 'name rank', ...] 並列印
- 修改 `main()` 以呼叫並使用函數 `extract_names()` 傳回的串列

** 先前我們習慣在函數中直接印出結果 (至螢幕)；但讓函數 *return* (傳回) 擷取的資料會使它更便於重複使用。因為呼叫者可選擇列印資料，也可用以做其它的處理。(當然你仍可同時在函數中直接列印以做為開發中的試驗)

6) 讓 `main()` 讀取使用者在命令列中指定的每個檔案，並呼叫函數 `extract_names()` 處理，以印出所擷取的文字總結。

** 可以使用 `join`：`text = '\n'.join(mylist) + '\n'` 將串列轉為更具可讀性的文字。

每個檔案印出的文字總結應類似以下格式：

```
C:> python childnames.py baby2006.html
2006
Aaliyah 91
Aaron 57
Abagail 895
Abbey 695
Abbie 650
...
```

Part B

[Part A]

不僅輸出結果到螢幕，我們也可將文字總結**儲存到檔案**。當使用者在命令列鍵入選項引數 `--summaryfile` 時，針對命令列中緊接在後的每個要解析的輸入檔案如：'foo.html'，我們將其解析結果寫入新的檔案名稱：'foo.html.summary'。

> 《未來發展》撰寫程式具備 `--summaryfile` 功能，當在命令列輸入："python childnames.py --summaryfile baby*.html" 時，便可用 * 對所有檔案執行程式，一個步驟就產生全部的總結檔案！(找出符合在命令列輸入的 "baby*.html" 樣式的所有檔案並建成串列，然後逐一讀取串列中檔名，執行 childnames.py。)

組織好爬取出的資料並存入 .summary 檔案之後，你便可隨時經由 PowerShell 命令列指令來檢索，如：

```
C:> select-string -path *.summary -pattern 'Alex '
baby2008.html.summary:75:Alex 85

C:> select-string -path *.summary -pattern 'Joy '
baby2008.html.summary:1007:Joy 548
```